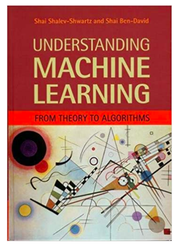
Certified Robustness Properties for Security Classifiers

COLUMBIA UNIVERSITY

Yizheng Chen, Shiqi Wang, Dongdong She, Suman Jana

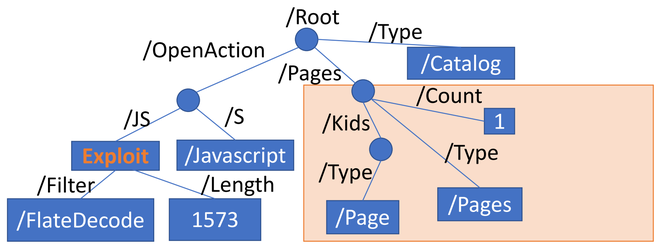
Evading Gmail's PDF Malware Classifier

Insert /Root/Pages from



to

Robustness Properties



Robust against building-block attack operations over malware. e.g., no matter how many pages are inserted.

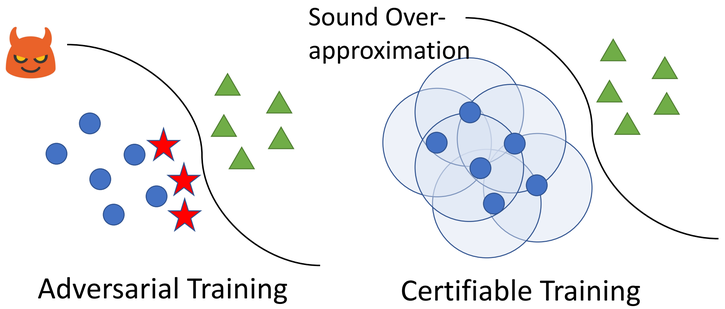
- **Subtree Insertion Property (Distance k)**

Robust against insertion attack in arbitrary k subtrees
e.g., k=1 includes /Pages, /Metadata, /OpenAction, etc.

- **Subtree Deletion Property (Distance k)**

Robust against deletion attack in arbitrary k subtrees
e.g., k=2 includes three choices here.

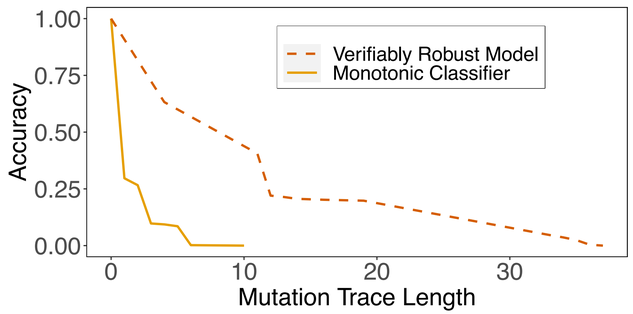
Certified Training to Learn Verified Robust Accuracy



- Measure property VRA: sound over-approximation to get VRA over a test set
- Multiple robustness properties: balance against different attackers
- Low false positive rate: small subtree distance
- Robust against unrestricted adaptive attackers: generalize robustness against building-block attacks to unbounded attacks

Key Results for Robust PDF Malware Classifier

- We achieve 92.27% average verified robust accuracy (VRA) over three properties, with 99.74% accuracy and 0.56% false positive rate.
- Our model has 7% higher robust accuracy against unrestricted whitebox attacks, than regular training and adversarial training.
- Our model requires 3.7 times more mutations and 10 times larger L0 distance to be evaded by adaptive attackers.



Twitter: @surrealyz
Code and models:
https://github.com/surrealyz/pdfclassifier